Root Cause Analysis for PBI000000000152:

Hardware issues with SL8500 robotic tape libraries in GCC

on 22 June, 2010

Report submitted 06 July, 2010

(CD-doc-4001)

Summary of Incidents

Frequent bot failures on SL8500 robotic tape libraries result in Enstore service disruptions potentially affecting CMS Tier 1 production data flow as well as data flow of other customers.

A failure of a single bot eventually leads to the outages in all three production Enstore systems with CMS experiment affected the most because of the large data flow and existing outage tolerance levels.

- CMS Tier 1 allows 10TB or 10,000 of CUSTODIAL files per queue at Fermilab that are not yet stored to tape. Upon tripping these thresholds the data from Tier 0 stops flowing automatically disrupting overall CMS production cycle. These thresholds can be reached easily withing 4-6 hours of peak data taking.
- Even if the thresholds are not reached, the data caching layer (dCache) creates additional load on Enstore by retrying to put files to tape resulting in huge queues both in Enstore and in dCache pools.
- User requests are queued in tape queues instead of going to pools.
- User requests for the same data that is not yet available on tape overwhelm the receiving pools that already suffer from huge tape queues (in terms of memory, CPU and file descriptor consumption).

It takes 3-4 hours of Enstore downtime to create such ripple effects across CMS Tier 1 dCache system.

List of related Incident tickets:

- INC000000039364
- INC000000039403

- INC000000027765
- INC000000030451
- INC000000031468
- INC000000033164
- INC000000036281
- INC000000039364
- INC000000039403
- INC000000039566

Background Concepts

The following points are useful in understanding the nature and impact of this problem.

Robotic tape libraries are accessed by Enstore which is a hierarchical mass storage system developed at Fermilab that provides seamless access to data by client applications distributed across IP networks.

Enstore is a client-server application. The server side is a multi-component structure of distributed servers that provides:

- hierarchical metadata view of user files stored on tape, presented to client as if it were a Unix file system
- storage/retrieval of individual files or group of files on tape in a manner which is close to simply copying file(s) between directories
- management of user files (e.g. renaming or removal)
- distributed access to tape drives
- interface to robotic tape libraries
- resource management of available tape drives
- tape allocation accounting per storage group, media type
- self-monitoring, error-reporting and alarm services

A simplified schematics of Enstore architecture is presented in Figure 1.

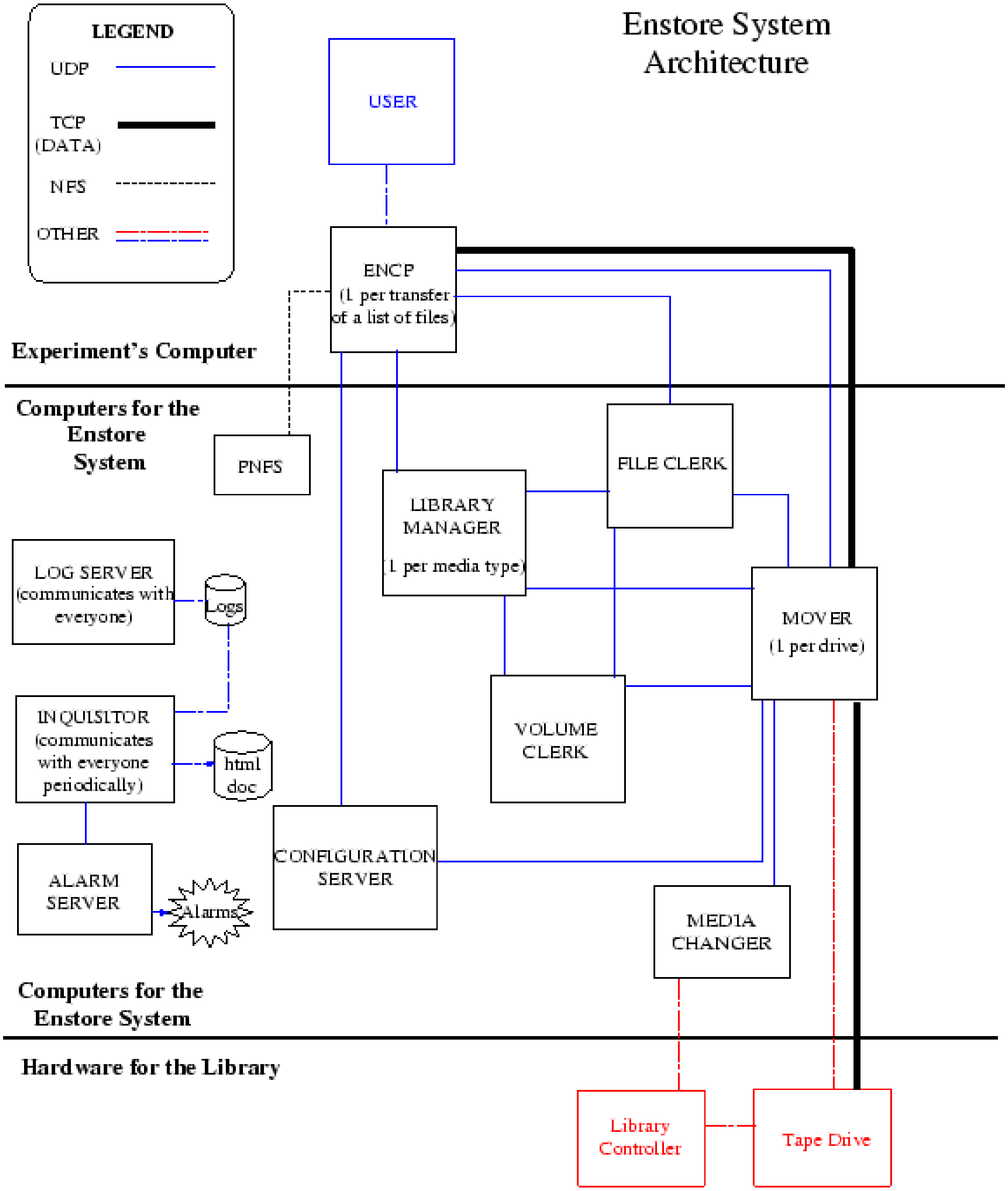

Enstore System Architecture
LEGEND
UDP
TCP (DATA)
NFS
OTHER
USER
ENCP (1 per transfer of a list of files)
Experiment's Computer
Computers for the Enstore System
PNFS
FILE CLERK
LIBRARY MANAGER (1 per media type)
LOG SERVER (communicates with everyone)
Logs
MOVER (1 per drive)
INQUISITOR (communicates with everyone periodically)
html doc
VOLUME CLERK
CONFIGURATION SERVER
ALARM SERVER
Alarms
MEDIA CHANGER
Computers for the Enstore System
Hardware for the Library
Library Controller
Tape Drive

Illustration 1: Simplified Enstore architecture

The problem being investigated here manifests itself in frequent Enstore downtimes (or outages) caused by intermittent hardware failures of HandBots (mechanical arms) in SL8500 robotic tape library, which is designated in the drawing as “Hardware for the Library”. The robotic tape library and its control software are provided by vendor – Oracle/SUN/StorageTeK.

A user of the system executes an encp (Enstore client application) on a computer, that has mounted pnfs namespace, to read or write a file (or list of files). Enstore components work in concert to identify a tape on which the file is available based on file ID, or to which the file can be written based on combination of destination directory pnfs tags. After the volume (tape) label has been identified, a Library Manager that manages this tape selects an idle Mover, which is a component of Enstore that runs on a computer that has tape drive attached to it. The Mover requests a Media Changer server to mount this volume in the Mover's tape drive. The Media Changer interacts with robotic tape library by executing Automated Cartridge System Library Software (ACSLS) commands via rsh on tape library control computer. Per these commands a mechanical HandBot picks the tape from its slot[1], transfers it to the location of tape drive which is attached to the Mover, and inserts the tape into the drive. The data then transferred directly from Mover to client machine (or vice versa) over TCP/IP.

Fermilab operates 3 different Enstore instances – d0en, cdfen and stken. The first two are used almost exclusively by D0 and CDF and the latter is shared between CMS and the rest of Fermilab experiments. A single tape library serves all three Enstore instances. Drives/tapes belonging to different Enstore instances and storage groups (experiments) are spread across all existing robotic tape libraries.

Within a given Enstore instance:

- A single Library Manager serves tapes of distinct media type and location (E.g. 1 LTO4 Library Manager in cdfen in FCC is "CDF-LTO4F1"). Library Manager spans tapes located in multiple robotic tape libraries connected together via Pass Through Ports (PTPs).
- Each Library Manager handles several Movers.
- One Media Changer serves all drives in a tape library unit.
- Media Changer is not aware of internal structure of robotic tape library (no concept of slot, rail, shelf or Bot).

Tape drives are not shared between Enstore instances.

Oracle/Sun/StorageTeK SL8500 is a top of the line enterprise-class modular robotic tape library. Currently Fermilab operates 5 such libraries with a 6th being commissioned. Each library has 10,000 slots

[1] There are also elevators involved besides the pass through ports to move a tape from its slot to the appropriate drive and vice. versa. Thus at GCC, as many as 3-4 bots, 2 pass thru ports, and an elevator may be involved in mounting a single tape. If its in the far left or right unit and has to move to a drive/slot in the opposite far right or left unit.

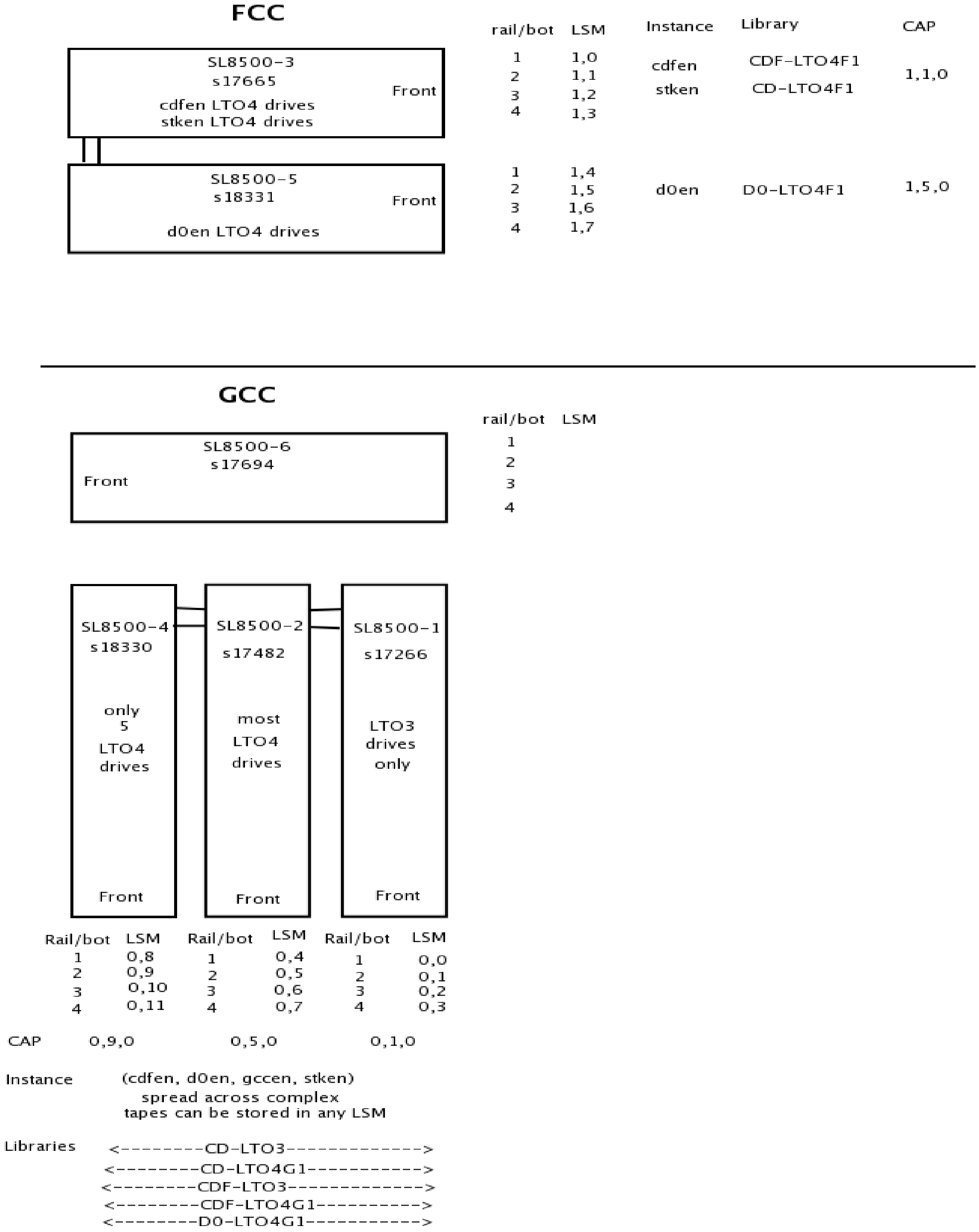


Illustration 2: Location and layout of existing production robotic tape libraries.

capacity. Location and layout of all existing SL8500s are shown in Figure 2. SL8500-3 and SL8500-5 are currently located in FCC and three SL8500s : SL8500-1, SL8500-2, SL8500-4 that are currently in operation are located in GCC. Sets of tape libraries in each building are connected into singe units by means of PTPs. Each SL8500 has 4 horizontal rails arranged from top to bottom. Each rail has moving HandBot serving a shelf of tape cartridges and a subset of tape drives.

A failure of a single HandBot on a rail leads to:

- failure of Media Changer to finish mount/dismount request;
- tape is set to NOACCESS
- Mover goes OFFLINE

Enstore has no knowledge of internal structure of the tape library and therefore does not know about bot failure or tapes locations inside tape library (hence doesn't know that a requested tape may be inaccessible).

As Enstore generates requests to tapes located in the affected shelf more and more tapes are marked NOACCESS and more movers become OFFLINE. When 50% of movers belonging to a given Library Manager becomes OFFLINE , the Enstore internal monitoring and alarm system creates so called “Red Ball” and generates high priority ticket in the Remedy System, which in turn pages SSA operator on call. The Service Desk incident INC000000039364 is an example of an automatically generated ticket. Upon receiving the page the following actions take place:

- Login to *ensrv4n, rsh from it to fntt node, check Automated Cartridge System System Administrator (ACSSSA) logs. Determine that this is a Bot failure. Takes few minutes (<10 min).
- Pause affected libraries. Send mail to customers.
- Open Service Desk ticket with request to place Oracle service call for a given CI (Configuration Item – ITIL speak).
- Service Desk calls back, initiates Oracle call.
- Service Desk calls back saying that they initiated Oracle call.
- Wait for Oracle to call back (~ 30 minutes).
- Wait for Oracle to arrive on site.
- Oracle expert replaces/repairs a broken HandBot.
- The affected Bot must sync it's firmware. Usually it is not at the correct level. The SL8500 has the firmware level and downloads to the bot. This usually takes about 20 minutes.
- Once the bot has the correct firmware, it runs its self diagnostics which takes another 5 minutes or so.
- When the access door is closed following the repairs, all the bots

that were not replaced already have the correct firmware so they immediately go through their self-diagnostics. They calibrate themselves by running back and forth the entire length of the rail and bang into the end stops (slowly). Once this happens they quickly run back and forth and stop before the ends are hit. They also scan each pass thru port, each elevator, and then each configured tape drive. Once they finish these checks, they are ready for service and begin to audit all the slots on their rails.

- Once the Bots begin their slot audits, the library managers can be released but not until SSA manually dismounts any tapes which failed to dismount due to the Bot outage.

- NOACCESS flag of all affected tape needs to be reset:

    - For the tapes that failed to mount, these can be cleared. (These can be cleared w/o waiting for the bot hardware fix).

    - For the tapes that failed to dismount, these need to be manually dismounted after the Bot has been replaced (and after it has started its slot audit).

- All movers offline/error/unknown states need to be restarted. Sometimes they need to be rebooted.

- Un-pause affected libraries.

- Send mail to affected customers noting when the libraries have been released.

- Additional work:

  The slot audits usually find a number of "upside down labels".

  The tapes are not upside down but for some reason the bots cannot read them. Upon completion of the slot audits, a list of the unreadable label tapes must be hand written and another procedure followed to have them ejected then reinserted.

  The Oracle tech does this work. Sometimes the SSA on-call has to remind them to make sure they do this. Otherwise we start to get new alarms as these tapes have transfers queued and the bots fail to read their physical labels, and the tapes go NOACCESS.

It is not until all 4 bots have started their slot audits that the libraries can be released. The slot audits are non-interfering thus tape mounts/dismounts take priority. When the bot finishes all the outstanding work, it picks up the audit where it left off. This process repeats until the audit has been completed.

SSA would like to get the libraries released ASAP; to get back in service/on-line to minimize production disruption. The Bots need to start processing mounts while the tech is on-site so we can be sure everything is working properly before they leave.

Up to 4-6 hours may elapse between getting “Red Ball” page and getting Oracle expert on site to replace a broken HandBot.

Timeline

Since the number of incidents is quite large we would follow up on one of the incidents based on records in Service Desk system. Incident INC000000039364:

- 06/06/2010 8:46:24 PM opened automatically by Enstore alarm service. Resulted in page to SSA operator on call. SSA primary must check Enstore web pages to see what system is affected.
    - Summary “ENSTORE BALL IS RED - Ticket Generated”. (Note that page tell the affected system.
    - Notes “1275875185.76', 1275875185.755214, 'stkensrv2n.fnal.gov', 14769, 'enstore', 'I (1)', 'Enstore_Up_Down', 'ENSTORE BALL IS RED - Ticket Generated', 'RedBall', 'STK Enstore', None, {'r_a': (('131.225.13.58', 44407), 2L, '131.225.13.58-44407-1275875185.655843-14769-182894084096'), 'text': {'Reason': ['Insufficient Movers for CD-LTO4G1.library_manager']}}]”
- 06/06/2010 8:47:45 PM SSA shifter acknowledged the page. Between this and next item the following activity takes place: Enstore web pages is checked for red balls and alarms on affected system; e-mail is checked for additional ticket details; SSA primary logins to the system remotely and analyses the logs; Entv is also used to see of number of movers in RED/ERROR state.
- 06/06/2010 9:31:05 PM SSA shifter responds with request to make service call to Oracle support for SL8500#2 (gcc center) - Prop # 101911 system number s17482. Bot failure.
    - Analysis of ACSSSA log indicates that the library errors were detected as early as of 06:02:16 PM.
- SSA shifter sends e-mail to Enstore customers that “Enstore outage of the GCC Libraries, suspected bot failure, d0en, stken, cdfen affected”

    - pauses CDF-LTO4G1, D0-LTO4G1, CD-LTO4G1 libraries.

- 06/06/2010 9:37:11 PM SSA shifter follows up with correct Serial Number and site ID.
- 06/06/2010 11:25:06 PM Service Desk sends service call request to Oracle via e-mail
- 06/06/2010 11:50:20 PM Service Desk sends another service call request to Oracle
- 06/07/2010 3:31:42 AM SSA shifter reports that Bot #4 in SL8500-2 (GCC) has been replaced.
    - Internal SSA log indicates that Bot #4 has been replaced at 2:50 AM.
    - 2:50 AM Additional 45 minutes downtime for restart/reboot the tape library system (ends about 3:32 AM)
- 06/07/2010 3:32:17 AM Resolution: Hardware was replaced by the Field Service Tech.
- 06/07/2010 3:32:17 AM Mail is sent to Enstore customers that the system is back in service.

There are slightly different actions for out of hours vs. in hours pages with respect to placing the service calls through the service desk. For in-hours bot issues, SSA opens h/w incidents. For out-of-hours, SSA calls x2345, explains the issue and waits for a return call. The Service Desk is not staffed 24x7 so they get paged and there is additional wait time for them to call back. There is an escalation procedure to follow if they do not return call within 30 minutes.

Analysis

In the time period 01/2009-07/2010 there have been 27 Bot failures requiring replacement. There is total 5x4=20 Bots, 8 in FCC and 12 in GCC. Figure 3 shows breakdown of bot failures by tape library and bot number. Immediately it can be seen that no bot failures have occurred in SL8500-4[2] which is located in GCC and has only 5 LTO4 tape drives (See Figure 2) installed recently and therefore is not used as heavily as for instance SL8500-2 which has the most number of failures – 13 and mostly operates LTO4 drives. Bots are numbered from 1 to 4 following from top to bottom rails. The Bot failures across tape libraries do not seem to be localized to one particular rail. There has been total 10 Bot

[2] During review of this document it was noted that top Bot #1 (LSM 0,8) was replaced on 06/08.

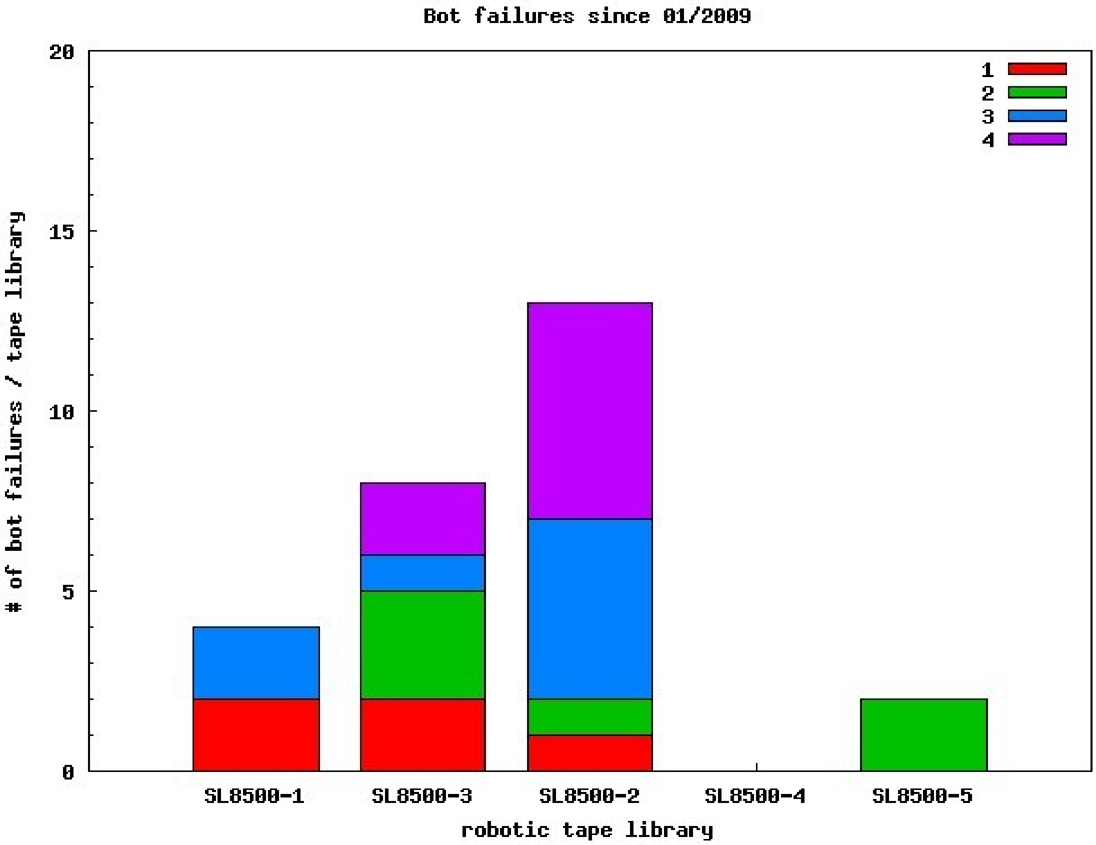
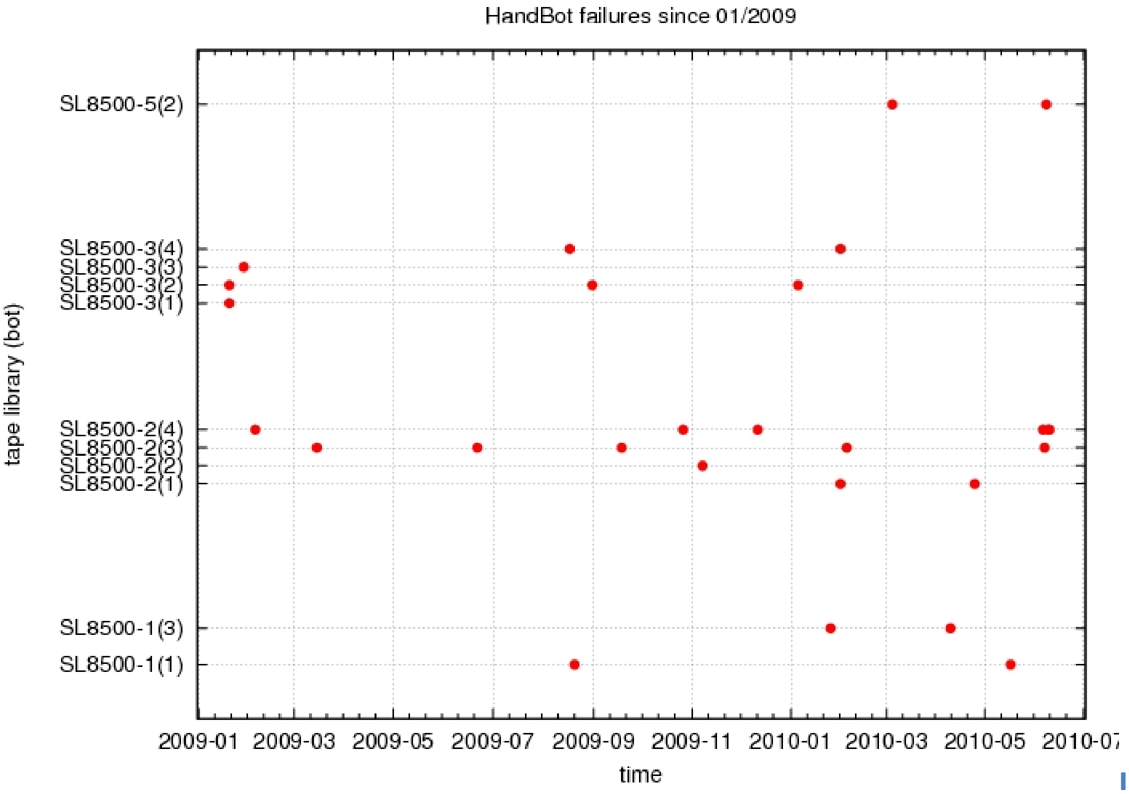


Illustration 3: Breakdown of failures by tape library and Bot number since January 2009

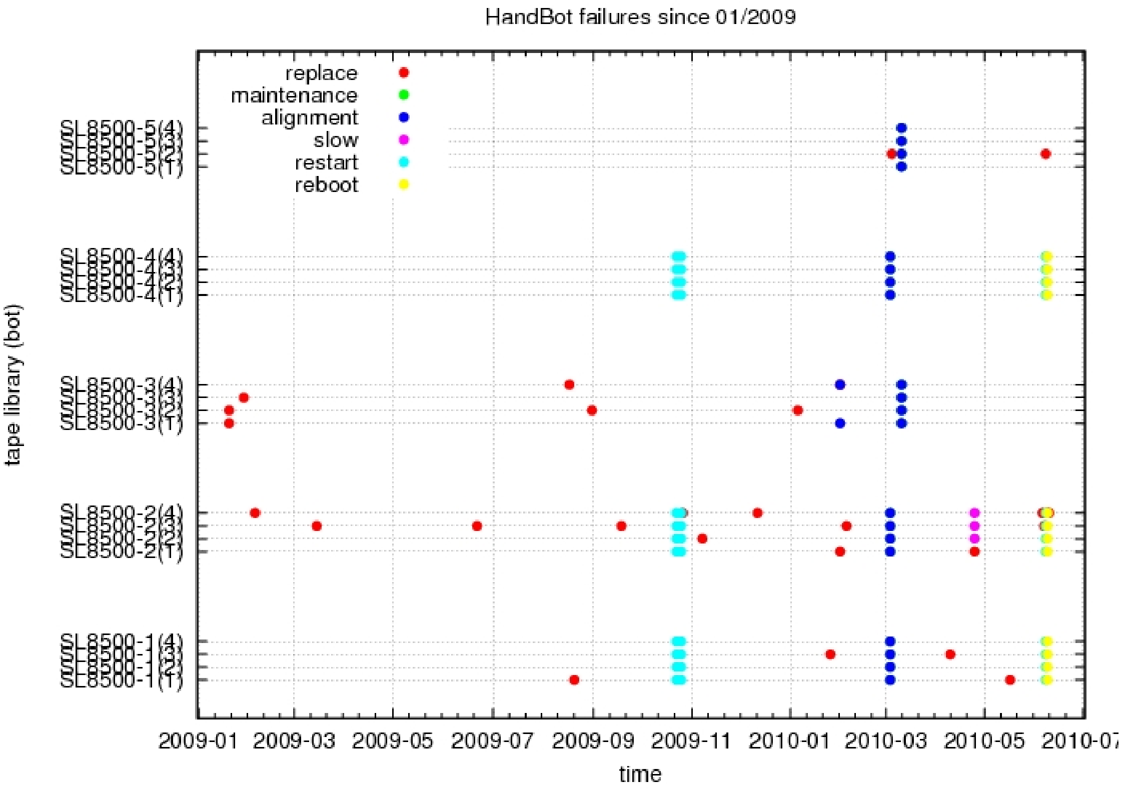
failures in FCC (SL8500-3 and SL8500-5) and 17 Bot failures in GCC (SL8500-1, SL8500-2 and SL8500-4).

An accounting database of Enstore system stores information about mounts/dismount operations performed by all movers in each production system. During last 19 months there has been 3849683 mount/dismount operations in GCC tape libraries (SL8500-1, SL8500-2 and SL8500-4) and 2407680 mount/dismount operations in FCC tape libraries (SL8500-3 and SL8500-5).

Based on these numbers we can calculate average number of mount/dismounts between the Bot failures observed. For FCC it is 2407680 / 12 / 17 = 11802 and for GCC it is 3849683 / 8 / 10 = 48121. These numbers need to be compared with expected mean exchanges/swaps between failures (MEBF/MSBF), which is advertised by Oracle to be 2M. Apparently we observe gross discrepancy between observed number of cartridge swaps (which we assume what mount/dismount operations are) and expected number. Even if we assume that 3-4 Bots are involved in mounting a single tape. This is a clear indication that either Bots or rail alignment or rack gear mechanisms are faulty. Oracle is currently performing analysis of one of the failed Bots with results expected soon (within a month?).



llustration 4: Timeline of Bot failures that rerquired replacements broken down by tape libraries for the time period since January 2009.



The main contributing causes to the problem denotes as unacceptably frequent SL8500 downtimes due to h/w Bot failure are depicted in fishbone diagram in Figure 6.

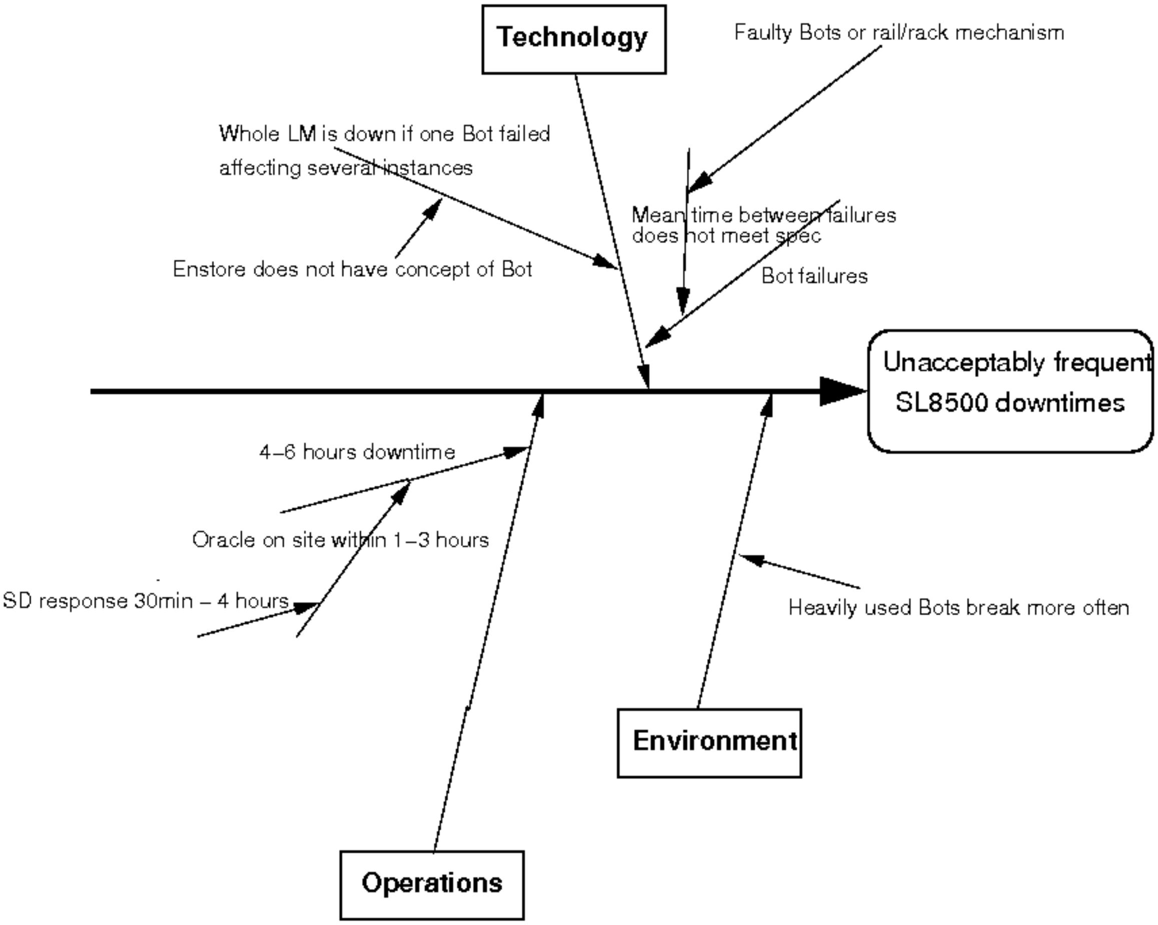

Technology
Faulty Bots or rail/rack mechanism
Whole LM is down if one Bot failed affecting several instances
Mean time between failures does not meet spec
Enstore does not have concept of Bot
Bot failures
Unacceptably frequent SL8500 downtimes
4–6 hours downtime
Oracle on site within 1–3 hours
SD response 30min – 4 hours
Heavily used Bots break more often
Environment
Operations

Illustration 6: Fishbone diagram of problem cause analysis

Ostensibly frequent Bot failures due to mechanical problems with either mechanical arm itself, rail alignment or rack gears is direct and root cause of long and frequent Enstore downtimes that require on-site visit by Oracle technicians to address the failures by replacing failed equipment.

Rack alignment does not seem to play a role as no significant improvement has been seen after rack alignment procedures (See Figure 5). Specifically the count of number of failures for the same time period since and before alignment shows improvement only in SL8500-3.

The problem is not localized to particular robotic library, or environment.

The problem is intermittent.

Direct Cause

Faulty HandBots or/and rack gear mechanism that do not meet specification in terms of mean time between failures are considered to be direct cause of the problem.

Contributing Factors

The following factors contribute to prolonged downtimes that affect multiple Enstore instances at Fermilab in case of failure of a single Bot in one of the SL8500 libraries. There is no significance to the ordering of these contributing factors.

- In many cases it takes inordinate amount of time between SSA responding to the page and Service Desk placing service call to Oracle technical support, sometimes up to 3 hours.

- MC does not differentiate errors caused by Bot failure. Therefore and alarm is not triggered immediately upon Bot failure. Instead the system continues to operate gradually degrading until 50% of movers attached to affected Library Managers go OFFLINE which may take hours depending on pattern and extend of user activity. For instance in case of INC000000039364 the time difference between first bot errors appeared in ACSSSA log file and automatic page was 2 hours 45 minutes.

- Library Manager straddles several robotic tape libraries. Failure of a single Bot in one of the tape libraries leads to eventual outage of Library Manager even if the rest of the system perfectly healthy. This is exacerbated by the lack of free slots. There have been cases where a Bot in SL8500-1 goes bad, which are mostly LTO3 tapes and only has LTO3 drives, causes LTO4 tapes to go NOACCESS as many are located in home slots in SL8500-1.

- Media Changer is not aware of Bot or rail and therefore will not avoid trying to mount tapes from affected rail.

- Lack of free slots in the SL8500-1 and SL8500-2 which forces tape mounts/dismounts to take longer due to their having to travel multiple bots, pass thru ports, elevators.

Root Cause

Faulty HandBots or/and rack gear mechanism that do not meet specification in terms of mean time between failures is considered to be a root cause of the problem. For example the rail bushing on a Bot continually running on mis-aligned racks may wear prematurely resulting in all kinds of symptoms, etc.

Recommendations

Based on the Problem investigation and root cause analysis of SL8500 hardware service disruption, the following list of recommendations were made for preventing future delays in timely notification and restoration of service disruptions to this system. These recommendations are very preliminary and no effort has been agreed to be committed to their implementation:

1. CMS is considering investing $130K into second redundant arm, one per side in tape library. This would allow to service broken Bot without service interruption. (Only for SL8500-{1,2,4})

2. Use many independent arms (e.g. 2 arms per rail). (All other libraries).

3. CMS is proposing to use other robots to be used in case of emergency. E.g. If Bot has failed in GCC tape library, redirect data flow to FCC tape library by dynamically modifying pnfs library tags Although it may cause other problems such as logistics of moving the tapes between buildings and will add more complexity to tape management.

4. Split robots, do not use Pass Through Ports. This will increase number of LMs to maintain, but will localize robot that contains failed Bot, so the effect on all Enstore production instances will be minimized. On the flip side this can lead to idle drives in one library while another has a backlog of requests.

5. Consider installing Oracle Service Delivery Platform (SDP) a part of Automated Service Request (ASR) system that will be generating service request with Oracle directly, by-passing Service Desk communication. This system needs to be certified by Computer Security.

6. Make Media Changer more intelligent. Media Changer should be aware of a rail on which a Bot has failed. Library Manager also needs to be modified. Estimated delivery time – 4 months. Need test facility.

7. Use Oracle ACSLS API library instead of rsh to communicate with SL8500 from Media Changer because Oracle basically ignores issues with non API access. Need test facility.

8. Drives need to be spread more uniformly between the tape libraries.

9. Need to improve monitoring and diagnostics of the Bot issue by providing mount count per rail (and per Media Changer and

media type).

10. Install the new racks and fasteners fix from Oracle. These may be the cause of at least some of the outages.

Root Cause Analysis Committee

The following people served on the RCA committee (22 June 2010):

Jon Bakken

John Hendry

Dmitry Litvintsev (lead)

Alexader Moibenko

Gene Oleynik

Darryl Wohlt

Michael Zalokar